4DFAB: A Large Scale 4D Facial Expression Database for Biometric Applications

Shiyang Cheng[1] Irene Kotsia[2] Maja Pantic[1] Stefanos Zafeiriou[1]

[1] Imperial College London [2] Middlesex University London

[1]{shiyang.cheng11, m.pantic, s.zafeiriou}@imperial.ac.uk [2]I.Kotsia@mdx.ac.uk

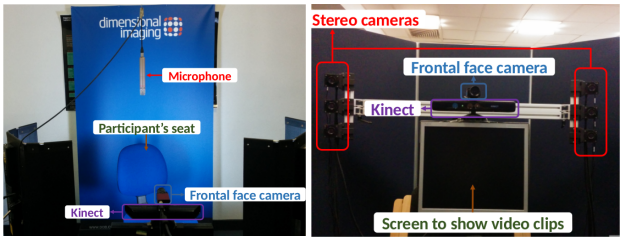
Abstract

The progress we are currently witnessing in many computer vision applications, including automatic face analysis, would not be made possible without tremendous efforts in collecting and annotating large scale visual databases. To this end, we propose 4DFAB, a new large scale database of dynamic high-resolution 3D faces (over 1,800,000 3D meshes). 4DFAB contains recordings of 180 subjects captured in four different sessions spanning over a five-year period. It contains 4D videos of subjects displaying both spontaneous and posed facial behaviours. The database can be used for both face and facial expression recognition, as well as behavioural biometrics. It can also be used to learn very powerful blendshapes for parametrising facial behaviour. In this paper, we conduct several experiments and demonstrate the usefulness of the database for various applications. The database will be made publicly available for research purposes.

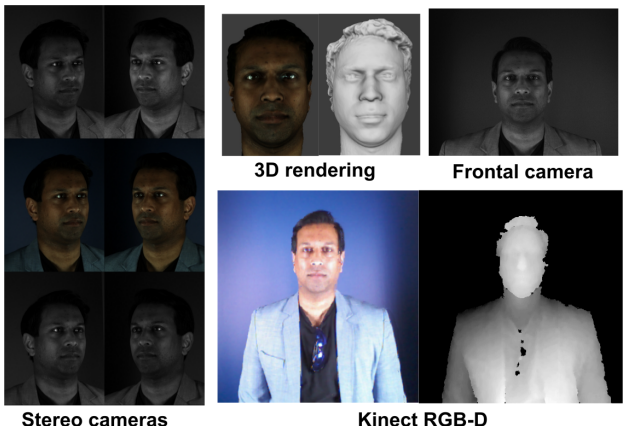
1. Introduction

In the past decade we have witnessed the rapid development of 3D sensors with which we can capture the facial surface (aka 3D faces). Immediately it was evident that a new stream of research could be opened and researchers started collecting databases of 3D face for many face analysis tasks, such as face recognition (FR) and facial expression recognition (FER).

BU-3DFE database [46], which includes articulated facial expressions from 100 adults, is probably the earliest and most popular database of expressive 3D faces. From then on, many static databases for 3D expression analysis [39, 33, 28, 26, 52, 40] were released and contributed largely to the development of fully automatic 3D FER systems. Similarly, half a decade ago, the releases of several 4D facial expression databases, BU-4DFE [45], D3DFACS [23], and Hi4D-ADSIP [32], expanded this line of research to the 4D domain. Nonetheless, the above emotion corpuses focused on only posed behaviours,



(a) Frontal (right) and rear (left) view of recording setup.



(b) Exemplar data from a single capture.

Figure 1: Overview of capturing system.

which hinders the use of 3D/4D FER system in real world scenarios. Henceforth, three databases, B3D(AC) [25], BP4D-Spontaneous [50] and BP4D+ [51], that captured dynamic 3D spontaneous behaviours were proposed. B3D(AC) dataset [25] is the first 4D audio-visual database, though its size (14 people) is small. Zhang *et al.* proposed BP4D-Spontaneous [50] database, which, not only tripled the subject number, but also introduced some well-designed tasks (*e.g.* interviews, physical activities) to elicit spontaneous emotions. BP4D+ [51] extended previous works by incorporating different modalities (i.e. thermal imaging, physiological signals) as well as more subjects. One common merit of BP4D-Spontaneous and BP4D+ is that they both provide expert FACS labels [24]

Name	Size (Ages)	Repeats	Content	FPS	Landmarks & Annotations
Chang *et al.* [20]	6 people (N/A)	N/A	6 posed expressions.	30	N/A
BU-4DFE [45]	101 people (18~45)	N/A	6 posed expressions.	25	83 facial points.
B3D(AC) [25]	14 people (21~53)	N/A	Spontaneous expressions and speech.	25	15 rated affective adjectives.
D3DFACS [23]	10 people (23~41)	N/A	Up to 38 AUs per subject.	60	47 facial points, AU peaks.
Hi4D-ADSIP [32]	80 people (18~60)	N/A	6 posed expressions + pain, face articulations and phrase reading with 3 intensities.	60	84 facial points.
Alashkar *et al.* [3, 4]	58 people (avg. 23)	N/A	Neutral and posed expressions with random poses, occlusion, talking, etc.	15	N/A
BP4D-Spontanous [50]	41 people (18~29)	N/A	Spontaneous expressions.	25	83 facial points, 27 AUs (2 with intensity).
BP4D+ [51]	140 people (18~66)	N/A	Multimodal spontaneous expressions.	25	83 facial points, 34 AUs (5 with intensity).
Ours	180 people (5~75)	4	6 posed expressions, spontaneous expression, 9 words utterances.	60	79 facial points.

Table 1: 4D facial expression databases. Size (Ages): Number of subjects and their age range. Repeats: Number of repeated sessions per subject. Content: Posed and spontaneous expression, etc. FPS: Frames captured per second. Landmarks & Annotations: Available landmarks and annotations.

which are very useful in emotion analysis. There are also some low resolution databases captured using the Kinect sensor [31, 6] designed for capturing 3D dynamic spontaneous behaviours.

Despite numerous 3D/4D facial expression databases are now publicly available, none of them contains samples collected in different sessions that allow us to investigate the use of dynamic behaviour for biometric applications[1]. As a consequence, research on dynamic 3D face recognition has fallen behind with static 3D face recognition. Only a few works were proposed in the past decade [41, 5, 27, 17], most of them used BU-4DFE database [45] which is limited for biometric applications. Arguably, the main reason is the lack of publicly available high quality 4D databases with many recording sessions that can be used for face recognition/verification. Furthermore, as the commonly used databases [3, 45] contain only one recording session per subject, the generalization ability of the tested method is doubtful.

As a matter of fact, all the aforementioned databases of 3D expressive samples (a) capture each subject only once (i.e. one recording session), which prohibits the use in a biometric scenario, (b) contain only posed or spontaneous expressions but not both and (c) generally include a small number of subjects.

In this work, we take a very significant step forward and propose the 4D Facial Behaviour Analysis for Security (4DFAB) database which includes 180 participants on 4 different recording sessions spanning a period of 5 years. 4DFAB database contains over 1.8 million high resolution 3D facial scans and has been collected from 2012 to 2017. We believe that 4DFAB is an invaluable asset for many different tasks such as 3D/4D face and facial expression recognition using posed/spontaneous behaviours, building high quality expressive blendshapes, as well as synthesizing 3D faces for training deep learning systems. To better compare the proposed database with existing 4D face databases, we give an overview in Table 1.

To summarise, our contributions are:

- We present a database (we refer as **4DFAB**) of 180 subjects collected over a period of 5 years under four different sessions, with over 1,800,000 3D faces.
- Our database contains both posed and spontaneous facial behaviours. The spontaneous behaviours were elicited by displaying stimuli that could elicit a variety of behaviours (*e.g.*, from smile and laughter to cries and confusion).
- We investigate, for the first time, the use of spontaneous 4D behaviour for biometric applications [2].
- We demonstrate that expression blendshapes learned from our database are much more powerful than the off-the-shelf blendshapes provided by FaceWarehouse [18].

2. Data Acquisition

For the past five years, we have collected a comprehensive dynamic facial expression database (4DFAB) that can be used for 3D face modeling, 3D face and expression recognition, etc. In this section, we will provide the details of this database.

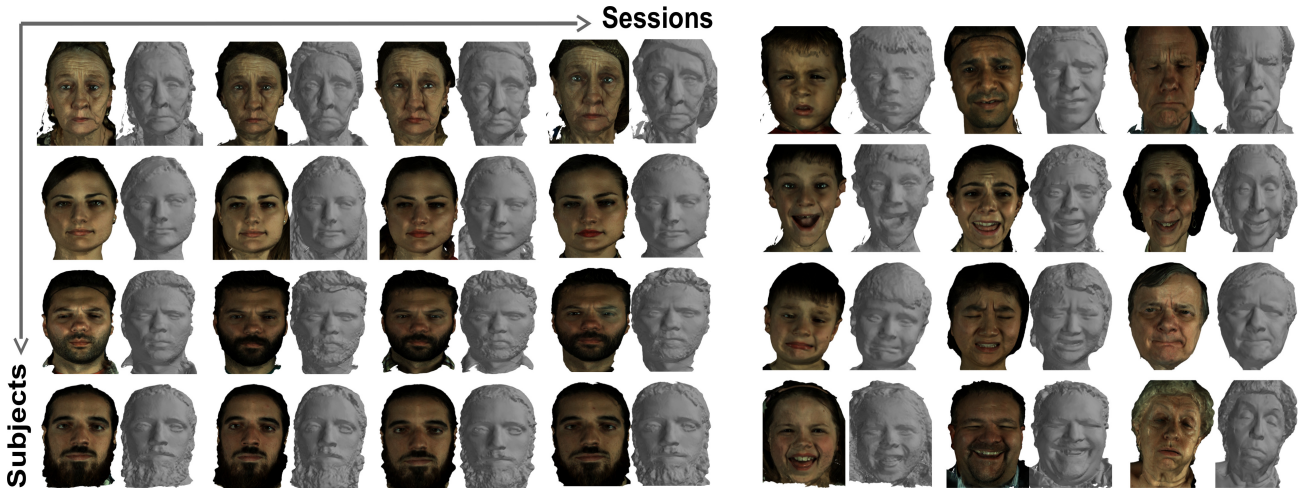
2.1. Capturing system setup

We used the DI4D dynamic capturing system [3] to capture and build 4D faces. This capturing system mainly consists of six cameras (two pairs of stereo cameras and one pair of texture cameras, 60FPS, 1200x1600). The distance between the subject and camera plane is 140cm.

[1] To the best of our knowledge, the databases collected for biometric applications contain only static 3D faces [38, 37, 47, 11], hence their use for analysis of facial motion in biometric application is limited.

[2] The study in [10] only studied posed speech related and speech unrelated facial behaviour for biometric applications.

[3] http://www.di3d.com



(a) Examples of subjects in four sessions.

(b) Examples of spontaneous expressions.

Figure 2: Examples of 3D faces in the database.

Calibration was performed before every recording session, using a 10x10 20mm checkerboard. Two 4-lamp fluorescent lights were placed on each side to provide consistent and uniform lights. The complete set up is shown in Figure 1. Additionally, we added one microphone to record audio signal, one frontal grayscale camera (60FPS, 640x480) to record frontal face image, and a Kinect to record RGB-D data (30FPS, 640x480). They were synchronized with the 4D recording using trigger from the DI4D capturing system, and will be publicly available after the initial release of high resolution 4D faces. Nevertheless, we mainly focus on the 4D data in this paper.

2.2. Experiment design and emotion elicitation

Our experiments aim at capturing posed expressions, spontaneous expressions and any other evident facial movements. We define posed expression as the participant deliberately making the expression that has the same semantic meaning as the target expression. Spontaneous expression, as its name suggests, is the natural and spontaneous emotion shown by the participant during the experiment. Video clip viewing was our main way to elicit such expressions. Except for expressions, we collected some facial movements that might not correspond to any emotions, examples of which included flaring nostrils, biting lip, raising eyebrow, etc.

Each participant was asked to read and sign a consent form, which allowed the use of data for research purposes. Before the experiment, participant was asked to take off any glasses, hat or scarf, and wear hairnet/hairband to prevent occlusion if necessary. After that, we calibrated the cameras, adjusted the seat, and made a preview capture. In order to acquire high-fidelity captures, participant was asked to avoid large body and head movements during the recording.

Within each recording session, we first asked the participant to perform 6 basic expressions (i.e. anger, disgust, fear, happiness, sadness and surprise) and pronounce the nine words (i.e. puppy, baby, mushroom, password, ice cream, bubble, Cardiff, bob, rope) three times in order. These two tasks were repeated for every session. Then, a few tasks involving words and numbers were undertaken. After this, we showed the participant several videos to elicit spontaneous expressions. Between two successive tasks, we gave the participant a short break (a minute or two) to reset his/her emotion state.

To provide 4D data for face recognition/verification purposes, we created 4 recording sessions with different video stimulus, and invited the same participant to attend 4 times. We list the tasks of each recording session of the proposed database in Table 2 3 4 5, which describe Session 1, 2, 3 and 4 respectively. Note that there might be multiple target emotions in one task, as we are anticipating different reactions from different subject. Examples of the same subject in different sessions are shown in Figure 2. We recorded over 40 hours of raw data, however, it was neither necessary nor feasible to reconstruct all of them. Therefore, we browsed every sequence and manually divided it into different segments of facial expression or movements. The final database included 1.8 million 3D meshes (equivalent of 8.4 hours recording), and took more than 20TB of storage space.

2.3. Participants

180 participants (60 females, 120 males) aged from 5 to 75 took part in the experiment. The majority of them

Task	Activity	Target Emotion
T1.1	Task: perform 6 expressions (anger, disgust, fear, happiness, sadness, surprise).	N/A
T1.2	Task: word utterances (puppy, baby, mushroom, password, ice cream, bubble, Cardiff, bob, rope), three times each.	N/A
T1.3	Video clip: jump scare immediately after words reading, finish with a joke.	Surprise, fear, happiness.
T1.4	Task: count backwards from 1000 by 7s.	Embarrassment, nervousness.
T1.5	Video clip: two disgusting and fearful clips on human eyes.	Disgust, fear, surprise.
T1.6	Video clip: comedy *Women: Know Your Limits!*	Happiness, anger.
T1.7	Video clip: funny moments of cat.	Happiness.
T1.8	Video clip: funny dance by a parrot.	Happiness.

Table 2: Tasks of Session 1 recording.

Task	Activity	Target Emotion
T2.1	Task: perform 6 basic expressions (same as T1.1).	N/A
T2.2	Task: word utterances (same as T1.2).	N/A
T2.3	Task: add 1 to each digit in the given numbers.	Embarrassment, nervousness.
T2.4	Task: add 3 to each digit in the given numbers.	Embarrassment, nervousness.
T2.5	Video clip: jump scare concealed in a relaxing video.	Surprise, fear.
T2.6	Video clip: a collection of disgusting but funny movie scenes.	Disgust, happiness.
T2.7	Video clip: emotional Thai story.	Sadness.
T2.8	Video clip: clip from the movie *Up*.	Sadness.
T2.9	Video clip: Jimmy Kimmel's *Halloween Candy Prank*.	Happiness.

Table 3: Tasks of Session 2 recording.

were recruited from our institute's administrative section and departments (Engineering, Business, Medicine, etc.), the other subjects (over 40) were volunteers from outside the college. They are from over 30 different culture backgrounds including Chinese, Greek, British, Spanish, etc. Ethnicity includes Caucasian (Europeans and Arabs), Asian (East-Asian and South-Asian) and Hispanic/Latino. The distribution of ages (based on the first attendance) and ethnic groups are summarised in Table 6.

Among all the participants, 179 subjects attended the first session, 100 subjects participated the second session, while 81 and 75 participants have come for the third and fourth time respectively. The average time interval between two consecutive attendances is 219 days (shortest: 1 day, longest: 1,654 days). Among them, 56% are recorded within 3 months, 23% are between 3 to 12 months and 21% for over a year.

2.4. Data processing and organization

Six synchronised 2D video sequences were recorded during experiment. For every pair of stereo images within the sequence, a passive stereo photogrammetry method was employed to produce a range map which was subsequently used for reconstructing 3D face. Ten machines were actively running for 1.5 years to reconstruct nearly two million selected frames. A summary of reconstructed data is given in Table 7. The vertex number of reconstructed 3D meshes ranges from 60k to 75k, with the maximum edge length allowed in mesh being 2mm.

3. Establishing 4D Dense Correspondences

It is very important to establish dense correspondences between every mesh and an universal template. There are two popular approaches for this, one is through non-rigid image registration in UV-space [35, 23], another is directly aligning the template to the target mesh (*e.g.* using NICP [7, 14, 21]). They both provide accurate correspondences, whereas UV-based approaches are more powerful and computationally efficient (refer to [14] for an in-depth comparison). Although some intricate face parts (*e.g.* interiors of nostrils, ears) are precluded from the UV map, it should not affect our data which do not exhibit those details. In this section, we explain our UV-space-based alignment framework (also demonstrated in Figure 3).

3.1. 2D to 3D mapping in UV space

Firstly, we create a 2D to 3D mapping by a bijective mapping from 2D positions in UV space to the corresponding 3D point in the mesh (see Figure 3(a)). Assume

Task	Activity	Target Emotion
T3.1	Task: perform 6 basic expressions (same as T1.1).	N/A
T3.2	Task: word utterances (same as T1.2).	N/A
T3.3	Task: subtract 1 to each digit in the given numbers.	Embarrassment, nervousness.
T3.4	Task: subtract 3 to each digit in the given numbers.	Embarrassment, nervousness.
T3.5	Video clip: jump scare concealed in a color blindness test video.	Surprise, fear.
T3.6	Video clip: man eating giant larva from *Man vs. Wild*.	Disgust.
T3.7	Video clip: Derek Redmond's emotional Olympic story in 1992.	Sadness.
T3.8	Video clip: funny fails compilation.	Happiness, surprise.
T3.9	Video clip: emotional shadow dance.	Sadness.

Table 4: Tasks of Session 3 recording.

Task	Activity	Target Emotion
T4.1	Task: perform 6 basic expressions (same as T1.1).	N/A
T4.2	Task: word utterances (same as T1.2).	N/A
T4.3	Video clip: jump scare immediately after words reading.	Surprise, fear.
T4.4	Task: replace the character with corresponding number in the given strings.	Embarrassment, nervousness.
T4.5	Video clip: disgusting but funny prank.	Disgust, happiness.
T4.6	Video clip: funny parody on iPhone.	Happiness.
T4.7	Video clip: edited commercial *Mind the Gap*.	Sadness.

Table 5: Tasks of Session 4 recording.

Age	Num.	Prop.
5-18	5	2.8%
19-29	115	63.9%
30-39	41	22.8%
40-49	9	5.0%
over 50	10	5.5%

(a) Age distribution

Ethnic	Num.	Prop.
Caucasian	101	56.1%
Asian	63	35.0%
Hispanic/ Latino	16	8.9%

(b) Ethnicity distribution

Table 6: Distribution of age and ethnicity in our database.

Session	Participants	Frames	Avg. frames
S1	179	768,290	4,292
S2	100	409,272	4,093
S3	81	345,921	4,271
S4	75	312,030	4,160
Total number of frames = 1,835,513			

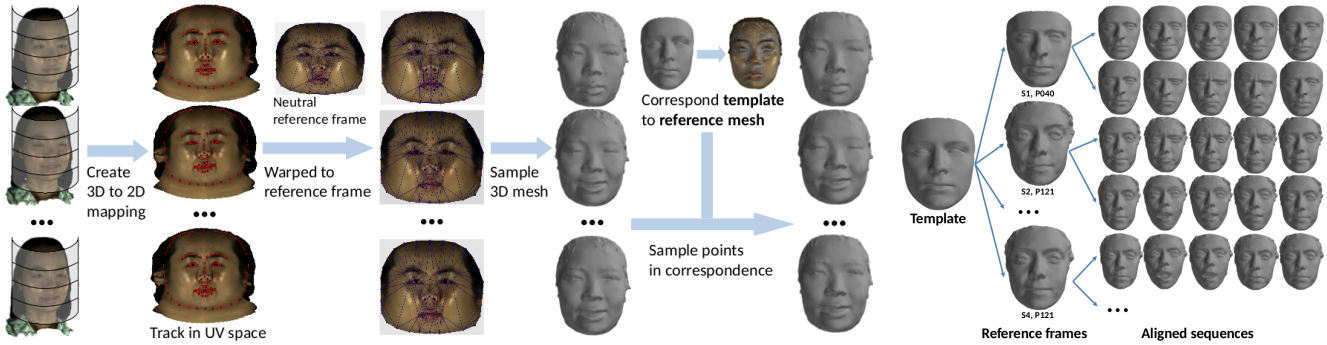
Table 7: Summary of reconstructed 4D data.

that such mapping could accurately represent a 3D face, establishing dense correspondence between any two UV images will automatically return us a dense 3D-to-3D correspondence for their corresponding 3D meshes. This is beneficial because it transfers the challenging 3D registration problem to the well-solved 2D non-rigid image alignment problem. Furthermore, in our specific case where 1.8 million meshes need to be aligned, this is obviously more reliable and computationally efficient. We employ an optimal cylindrical projection method [15] to synthetically create a UV space for each mesh, and produce a UV map I with each pixel encoding both spatial information (X, Y, Z) and texture information (R, G, B), on which we perform non-rigid alignment.

3.2. Non-rigid UV image alignment and sampling

Several non-rigid alignment methods in UV space have been proposed. One way (denoted as UV-OF) is applying Optical Flow on the UV texture and the 3D cylindrical coordinates to align two UV maps [12, 16]. Another approach utilises key landmarks fitting and Thin Plate Spline (TPS) warping [35, 23] (referred as UV-TPS). We follow the UV-TPS approach because UV-OF might produce drift artifacts as the optical flow tracking continues, while 2D landmarks detection usually provides stable and consistent facial landmarks to avoid drifting.

Compared with [23], we made several changes to suit our data. Firstly, we built session-and-person-specific Active Appearance Models (AAMs) [2, 8] to automatically track feature points in the UV sequences. This means that 4 different AAMs would be built and used separately for one subject. Main reasons behind this are (1) textures of different sessions differ due to several facts (i.e. aging, beard, make-ups, experiment lighting condition, etc.); (2) person-specific model is proven more accurate and ro-



(a) 4D alignment pipeline.

(b) Overview of correspondences.

Figure 3: Our framework for establishing 4D dense correspondences.

Figure 4: 4D sequences that are in full correspondence with template, displayed with 79 landmarks. For each sequence, top row shows original scans, middle and bottom rows are the registered 3D meshes with and without texture respectively.

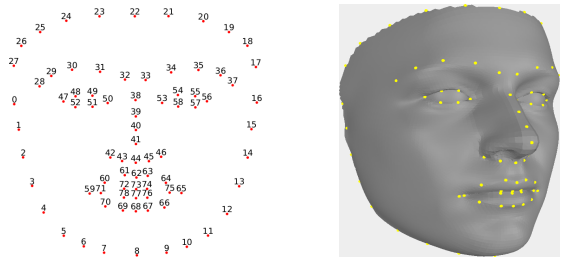


Figure 5: Definition of 79 facial landmarks (left) and exemplar of annotated template (right).

bust in specific domain [22]. Therefore, we manually selected 1 neutral and 2-3 expressive meshes per person and session, and annotated 79 3D landmarks (Figure 5) using a 3D landmarking tool [4] [1]. Overall, 435 neutral meshes and 1047 expression meshes were labelled. They were unwrapped and rasterised to UV space, and then grouped for building the corresponding AAMs. Note that we flipped each UV map to increase fitting robustness.

Once all the UV sequences were tracked with 79 landmarks, they were then warped to the corresponding reference frame using TPS, and thus achieving the 3D dense correspondence. For each subject and session, we built one specific reference coordinate frame from his/her neutral UV map. From each warped frame, we could uniformly sample the texture and 3D coordinates. Eventually, a set of non-rigidly corresponded 3D meshes under the same topology and density were obtained. Here, an extra rigid alignment step might be performed to further remove similarity differences.

[4] https://www.landmarker.io

3.3. Establishing correspondence to face template

Given that meshes have been aligned to their designated reference frame, the last step is to establish dense 3D-to-3D correspondences between those reference frames and a 3D template face. This is a 3D mesh registration problem, and can be solved by Non-rigid ICP [7]. NICP extends the ICP algorithm by assuming local affine transformation for each vertex, and iteratively minimises the distance between source and target meshes with adjustable stiffness constraint. We employed it to register the neutral reference meshes to a common template - Basel mean face [36]. We did not use the full Basel face, because our meshes might not fully describe ears, neck and nostrils. Thus we crop the original face and flatten the nostrils to get a new template (see Figure 5 for the modified template). Upon completion of this step, we corresponded every 3D mesh to one single template, an overview of correspondences is depicted in Figure 3(b). We also plot many example sequences in Figure 4.

4. Building Expression Blendshape Model

In order to build the blendshapes we used the methodology proposed in [34]. In particular, we annotated the apex frames (frames with maximum facial change) of all the pose expression sequences (anger, disgust, happiness, fear, sadness and surprise). For each of the sequences we subtracted the neutral mesh of the sequence from the corresponding apex frame. In this manner, we created a set of m difference vectors $\mathbf{d}_i \in \mathbb{R}^{3n}$ which were then stacked into a matrix $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_m] \in \mathbb{R}^{3n \times m}$, where n is number of vertices in our mesh. Afterwards a variant of sparse Principal Component Analysis (PCA) was applied to our data matrix $\mathbf{M}$ to identify sparse deformation components (the interested reader may refer to [34] for more details). Note that for visualization purposes, we blended all our meshes to the template provided by [14] to recover a face with ears and neck. We will provide both quantitative and qualitative measures to evaluate our expression blendshape in the next section.

5. Experiments

We conducted several experiments using the densely corresponded sequences, and report the performances as the baselines of our 4DFAB database for the research community.

5.1. Facial expression recognition

We did two standard FER experiments on 6 posed expressions, one was static and another was dynamic. Considering that not all the sessions had full attendance, separate subject-independent recognition experiments were set up. Within each session, we created a 10-fold partition, every time one fold was used for testing, the others were used for training. Both static and dynamic experiments used the same 10-fold partition.

Method	S1	S2	S3	S4
Ours	72.69	68.88	71.1	70.09
FW [18]	71.02	68.03	68.35	66.81

Table 8: Recognition Rates (RR) [%] obtained from 3D Dynamic facial expression recognition experiments.

5.1.1 3D Static expression recognition

We manually labelled the apex frames of each expression sequence. Because the apex interval varied from sequences, we trimmed the apex period and generated a balanced set that had 5 meshes per subject per session. We rasterised every 3D mesh into Z-buffer, and extracted the main face regions (covering eyes, mouth, cheeks and nose) based on 79 facial landmarks. The region was further divided into non-overlapping blocks, for which Histogram of Oriented Normal Vectors (HONV) [43] were computed. After this, PCA and LDA were used for dimensionality reduction, a multi-class SVM [19] was employed to classify expressions. Radial Basis Function (RBF) kernel was selected, whose parameters were chosen by an empirical grid search. We achieved a recognition rate of 70.27% Session 1 experiment, 69.02%, 66.91% and 68.89% in Session 2, 3 and 4 experiments respectively.

5.1.2 3D Dynamic expression recognition

We used Long Short-term Memory (LSTM) [29] to recognise dynamic expressions. For every expressive 3D face, we computed its facial deformation with regard to the corresponding neutral face. We then projected it to our blendshape model as well as the FaceWarehouse (FW) model [18] to obtain the sparse representations of expression, which would be used as the feature. In order to reduce noises, Kalman filter [9] was further applied to each dimension of features within the segments. For each experiment, only one standard LSTM layer was utilised, whose capacity was empirically decided according to the number of available training data. The Adam algorithm [30] was selected, with a learning rate of 0.001, batch size 12, and 15 epochs at max. Results in Table 8(b) showed that even with the simplest feature (i.e. blendshape parameters) and a basic LSTM network, we could achieve around 70% in recognition rate, which suggested that our 4D alignment was quite accurate and reliable. Moreover, recognition performances of our blendshape were better than FW in every session, which showed that

Figure 6: Comparison of FaceWarehouse blendshape model [18] and our expression blendshape model. Note that we only transfer the expression to the mean face of [14], not involving any modeling of face identity.

our blendshape could model expression more accurately.

5.2. 3D Dynamic face recognition

We report the results for 3D dynamic face recognition using 6 basic expressions respectively. We selected 74 subjects who have attended four sessions and performed all expressions. For each experiment (namely each expression), we performed a leave-one-out cross-validation - each time one session was left out for testing. The same feature, temporal filtering and LSTM network from Section 5.1.2 were employed for these tasks, except that LSTM capacity was decided experimentally. Both our expression blendshape model and FW were tested. Although our blendshape (named as **3DMM-exp**) models only the facial deformation and does not involve shape information of identity, it is essentially a variant of 3DMM. Therefore, it is interesting to compare it the standard 3DMM that models both expression and identity information. We built a 3DMM using 1482 aligned meshes with ground-truth 79 landmarks, and projected all the meshes to this model to obtain the shape parameters. We empirically selected the first 68 components and used it as feature descriptor (denoted as **3DMM** in Table 9).

From Table 9, not surprisingly, using 3DMM we could recognise 96% of the test instances on average. Since 3DMM jointly models the identity and expression, its shape parameters embed information from both sides, thus lowers the difficulty for LSTM to recognise test subject. Nevertheless, with our expression blendshape model in which identity is not present, we could achieve meaningful performances using anger, disgust and happiness (66.22%, 66.55% and 67.23% respectively). To the best of our knowledge, we are the first one to exploit dense shape deformation (without explicit identity information) in 3D dynamic face recognition. Our results also indicate that the use of dynamic expression sequences in a biometric scenario is worth investigating.

Method	AN	DI	FE	HA	SA	SU
FW [18]	45.61	51.69	41.89	54.73	45.95	49.66
3DMM-exp	66.22	66.55	62.84	67.23	59.46	61.15
3DMM	96.62	95.95	96.28	97.3	96.62	95.61

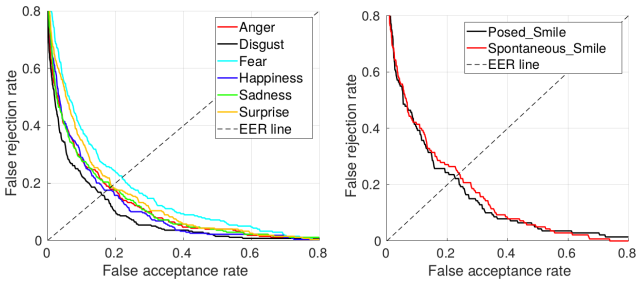
Table 9: Recognition Rates (RR) [%] obtained from 3D Dynamic face recognition experiments using 6 expressions (AN-Anger, DI-Disgust, FE-Fear, HA-Happiness, SA-Sadness, SU-Surprise).

5.3. 3D Face verification

Two verification experiments with posed expressions and spontaneous smile respectively were undertaken. We borrowed the verification methods from [48], in which the facial deformation was calculated between the neutral and expression apex frame, and used for verification. Supervised and unsupervised dimensionality reduction techniques were applied to extract sparse feature. In our cases, deformation was computed as the difference between aligned expressive mesh and its neutral mesh.

5.3.1 Verification using posed expressions

In this experiment, each posed expression was tested separately. Based on 4 recording sessions, 4 experimental sessions were implemented by employing the leave-one-out scheme and rotation estimates. For all the experiments, we excluded 131 subjects who did not attend all sessions and used them as test impostor claims. For each experiment, the leave-out session would be used as the test set (genuine claims), thus the other three sessions were used for training. The number of genuine



(a) Posed expressions (b) Spontaneous/posed smile

Figure 7: The ROC curves for (a) Posed expressions and (b) Spontaneous and posed smile.

claims was 70. We used only one apex frame per client. The rest training and testing procedures were identical as those described in [48]. We showed the verification performance in the Receiver Operating Characteristic (ROC) curve with False Rejection Rate (FRR) and False Acceptance Rate (FAR). The performance of a verification system is often quoted by a particular operating point of the ROC curve where FAR = FRR, which is called Equal Error Rate (EER) [49].

The curves are plotted in Figure 7(a), EER for anger, disgust, fear, happiness, sadness, surprise are 18.2%, 15.9%, 21.9%, 17.5%, 18.5% and 18.8% respectively. It suggests that, for our posed expression data, disgust (15.9%) and happiness (17.5%) are more informative than the others in verification.

5.3.2 Verification using spontaneous smile

To demonstrate the usefulness of our spontaneous data, we used the apex frame of spontaneous smile/laughter per subject and session for verification. The protocol was similar to previous verification experiment. Four sessions were implemented with the leave-one-out scheme. For the impostors, we reserved 124 subjects who did not have a full set of smiles from all sessions. The number of genuine clients across all sessions was 39. We also applied the same protocol for the posed Happiness to compare. The ROC curves are plotted in Figure 7(b). The EER achieved by posed smile was 22.6%, while spontaneous smile was 23.9%. The attained results indicat that spontaneous smile is as useful as its posed counterpart for automatic person verification. As far as we know, this is the first investigation on the use of 4D spontaneous behaviours in biometric application.

5.4. 3D Dynamic speech recognition

We conducted a dynamic speech recognition experiment on nine words: puppy, baby, mushroom, password, ice cream, bubble, Cardiff, bob, rope. This experiment

Session	S1	S2	S3	S4
LSTM	400	350	300	300
RR[%]	77.89	75.17	70.28	68.47

Table 10: Recognition Rates (RR) obtained from 3D Dynamic speech recognition experiments on 9 words utterances. LSTM capacity for each session was also reported.

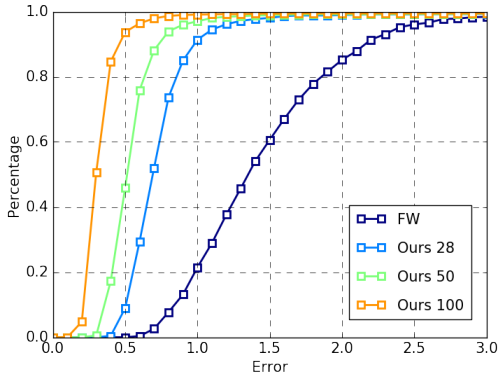
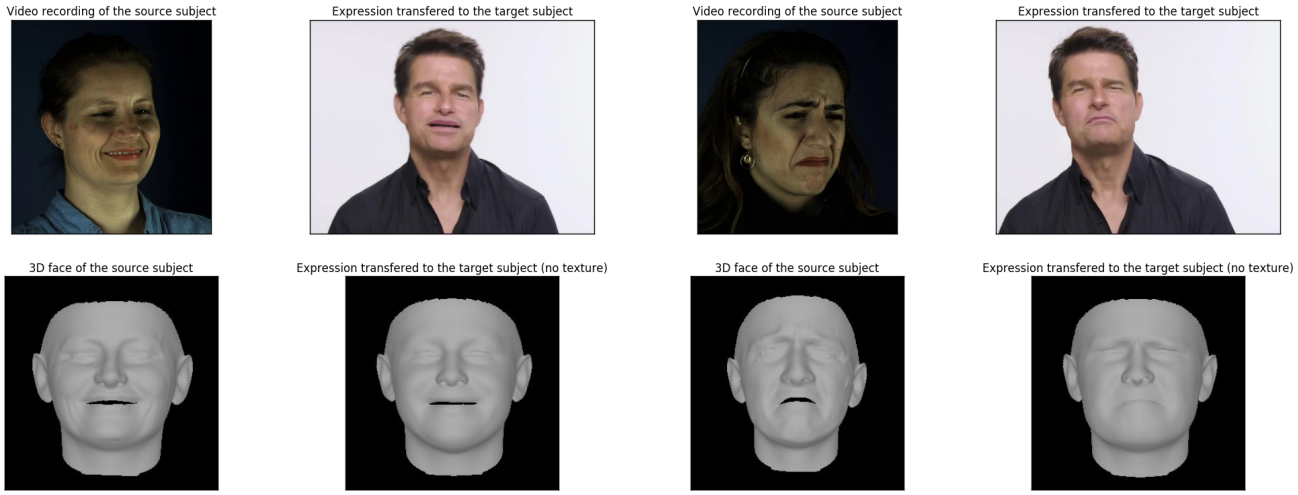


Figure 8: Cumulative reconstruction errors achieved with different blendshapes over randomly selected spontaneous expressions from our database.

was performed session-wise due to the same reason as expression recognition. For each session, we created a 10-fold partition, with one fold for testing, and the rest nine folds for training. The same LSTM [29] network was utilised. Since we were only interested in the mouth part, we defined a mouth region in the face template and extract it from every mesh in the sequence. Similarly, we built a 3D mouth model using all the meshes with landmark annotation, and kept 98% of the variations (46 components). All the meshes were then projected to this 3D mouth model to retain the shape parameters, which would be used as our feature descriptors. Kalman filter [9] was applied to smooth the feature sequence. As shown in Table 10, without any usage of texture or elaborated features, even our worse recognition performance (S4) is nearly 69%, while the best performance (S1) is 77.89%. This is quite likely contributing to an accurate dense alignment in mouth part.

5.5. Evaluation of expression blendshape model

We compare our blendshape model with FaceWarehouse (FW) model [18] in spontaneous expressions reconstruction. We randomly selected 1453 frames that display spontaneous expressions, from which, we computed the facial deformation in the same way as described in 4 and reconstructed it using both blendshape models. We calculate the reconstruction error and plot the cumulative error curves for models with different number



(a) Expression transfer given a spontaneous expression sequence from subject P085 of Session 4.

(b) Expression transfer given a spontaneous expression sequence from subject P163 of Session 4.

Figure 9: Screenshots of expression transfer demo video.

of expression components in Figure 8. To provide a fair comparison with FW, we report the performance of our model (**Ours-28**) using the same number of components as FW. Similarly, **Ours-50** and **Ours-100** denote model with 50 and 100 components respectively. It is clear that our blendshape model largely outperforms FW model, regardless of the number of components to utilise.

Additionally, we plot a dozen of 3D expression transfer examples in Figure 6. Specifically, we calculated the facial deformation of each unseen expression and reconstruct using our blendshape and FW respectively. We then cast the reconstructed expression on a mean face [14] for visualisation. Note that we fixed the number of our expression components to be identical with FW. It is obvious that our blendshape model can faithfully reconstruct unseen expressive faces with correct expression meaning. Moreover, our recovered shapes contain more facial details, such as wrinkles between the eyebrows.

5.6. Expression transfer

Our expression blendshape model, together with the dynamic sequence, can be used to synthesize new motion sequences from one single 3D face, with or without texture. In particular, we reconstruct each expression of the dynamic sequence from the source actor using our blendshape model, and transfer these expressions to the target actor. We demonstrate this application via an exemplar video of Tom Cruise downloaded from Youtube. There are three processing steps involved: (1) we perform a 3D Morphable Model (3DMM) fitting [13] on the given video, and select one 3D neutral face $\mathbf{T}$ among all the fitted meshes which will be our transfer target; (2) we select a 4D sequence $S = \{\mathbf{S}_1, ..., \mathbf{S}_n\}$ that exhibits rich expressions from one subject in our database, compute the dense deformations with regard to the subject's neutral face and reconstruct all the expressions using our blendshape model; (3) finally we can apply the reconstructed expressions $\Delta s = \{\Delta\mathbf{s}_1, ..., \Delta\mathbf{s}_n\}$ to the target $\mathbf{T}$ separately and obtain a new expressive sequence of Tom Cruise $T_{new} = \{\mathbf{T} + \Delta\mathbf{s}_1, ..., \mathbf{T} + \Delta\mathbf{s}_n\}$. For every newly generated mesh, we use the same UV texture from $\mathbf{T}$. There are quite a few elaborated methods to generate more realistic results, interested reader may refer to [42, 44]. For entertainment purposes, we rewind the target video to the same length of source sequence, and render every synthetic meshes to the extended video. We select two dynamic sequences and perform the expression transfer separately, Figure 9 shows the screenshot of each demo video. Note that our method can be applied to any given 3D faces, provided that it has the same mesh topology as our model.

6. Conclusion

We have presented 4DFAB, the first large scale 4D facial expression database that contains both posed and spontaneous expressions, and can be used for biometric application as well as facial expression analysis. We demonstrate the usefulness of the database in a series of recognition and verification experiments. We investigate, for the first time, the use of identity-free dense shape deformation from posed/spontaneous expression sequences in biometric applications. Promising results are obtained with basic features and standard classifier, thus we believe that dynamic facial behaviours could be further exploited

for face recognition and verification. Last but not the least, we build a powerful expression blendshape model from this database, which outperforms the state-of-the-art blendshape model. The database will be made publicly available for research purposes.

References

[1] J. Alabort-i-medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. ACM MM, pages 679–682, Orlando, Florida, USA, November 2014. 6

[2] J. Alabort-i-medina and S. Zafeiriou. A unified framework for compositional fitting of active appearance models. *IJCV*, 2016. 5

[3] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti. A 3D Dynamic Database for Unconstrained Face Recognition. In *5th International Conference and Exhibition on 3D Body Scanning Technologies*, Lugano, Switzerland, Oct. 2014. 2

[4] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti. A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions. In *ECCV Workshops*, Zurich, Switzerland, Sept. 2014. 2

[5] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti. A grassmann framework for 4d facial shape analysis. *PR*, 57(C):21–30, Sept. 2016. 2

[6] S. Aly, A. Trubanova, L. Abbott, S. White, and A. Youssef. Vt-kfer: A kinect-based rgbd+time dataset for spontaneous and non-spontaneous facial expression recognition. In *ICB*, pages 90–97, May 2015. 2

[7] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, pages 1–8, June 2007. 4, 7

[8] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *T-PAMI*, 37(6):1312–1320, June 2015. 5

[9] T. Basar. *A New Approach to Linear Filtering and Prediction Problems*, pages 167–179. Wiley-IEEE Press, 2001. 7, 9

[10] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall. Assessing the uniqueness and permanence of facial actions for use in biometric applications. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(3):449–460, 2010. 2

[11] C. Beumier and M. Acheroy. Face verification from 3d and grey level clues. *PR*, 22(12):1321 – 1329, 2001. Selected Papers from the 11th Portuguese Conference on Pattern Recognition - RECPAD2000. 2

[12] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. SIGGRAPH, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 5

[13] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild". In *CVPR*, 2017. 10

[14] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *IJCV*, Apr 2017. 4, 7, 8, 10

[15] J. Booth and S. Zafeiriou. Optimal uv spaces for facial morphable model construction. In *ICIP*, September 2014. 5

[16] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. *ACM TOG*, 29(4):41:1–41:10, July 2010. 5

[17] S. Canavan, X. Zhang, L. Yin, and Y. Zhang. 3d face sketch modeling and assessment for component based face recognition. In *IJCB*, pages 1–6, Oct 2011. 2

[18] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, Mar. 2014. 2, 7, 8, 9

[19] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM TIST*, 2(3):27:1–27:27, May 2011. 7

[20] Y. Chang, M. Vieira, M. Turk, and L. Velho. Automatic 3d facial expression analysis in videos. In *AMFG*, pages 293–307, Berlin, Heidelberg, 2005. Springer-Verlag. 2

[21] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic. Statistical non-rigid icp algorithm and its application to 3d face alignment. *IVC*, 58(Supplement C):3 – 12, 2017. 4

[22] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE SMCB*, 42(4):1006–1016, Aug 2012. 5

[23] D. Cosker, E. Krumhuber, and A. Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *ICCV*, pages 2296–2303, Washington, DC, USA, 2011. 1, 2, 4, 5

[24] P. Ekman and W. Friesen. *Facial Action Coding System*. Number v. 1. Consulting Psychologists Press, 1978. 1

[25] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool. A 3-d audio-visual corpus of affective communication. *IEEE T-MM*, 12(6):591–598, Oct 2010. 1, 2

[26] S. Gupta, M. K. Markey, and A. C. Bovik. Anthropometric 3d face recognition. *IJCV*, 90(3):331–349, Dec 2010. 1

[27] M. Hayat, M. Bennamoun, and A. A. El-Sallam. Fully automatic face recognition from 3d videos. In *ICPR*, pages 1415–1418, Nov 2012. 2

[28] T. Heseltine, N. Pears, and J. Austin. Three-dimensional face recognition using combinations of surface feature map subspace components. *IVC*, 26(3):382 – 396, 2008. 1

[29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computing*, 9(8):1735–1780, Nov. 1997. 7, 9

[30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 7

[31] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. Riek. 3d corpus of spontaneous complex mental states. In *ACII*, 2011. 2

[32] B. J. Matuszewski, W. Quan, L.-K. Shark, A. S. Mcloughlin, C. E. Lightbody, H. C. A. Emsley, and C. L. Watkins. Hi4d-adsip 3-d dynamic facial articulation database. *IVC*, 30(10):713–727, Oct. 2012. 1, 2

[33] A. B. Moreno and A. Sánchez. GavabDB: a 3D Face Database. In *Workshop on Biometrics on the Internet*, pages 77–85, Vigo, Mar. 2004. 1

[34] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and C. Theobalt. Sparse localized deformation components. *ACM TOG*, 32(6):179:1–179:10, Nov. 2013. 7

[35] A. Patel and W. A. P. Smith. 3d morphable face models revisited. In *CVPR*, pages 1327–1334, 2009. 4, 5

[36] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, Washington, DC, USA, 2009. 7

[37] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR*, volume 1, pages 947–954 vol. 1, June 2005. 2

[38] W. Sankowski, P. S. Nowak, and P. Krotewicz. Multimodal biometric database dmcsv1 of 3d face and hand scans. In *MIXDES*, pages 93–97, June 2015. 2

[39] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. *Bosphorus Database for 3D Face Analysis*, pages 47–56. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. 1

[40] G. Stratou, A. Ghosh, P. Debevec, and L. P. Morency. Effect of illumination on automatic expression recognition: A novel 3d relightable facial database. In *FG*, pages 611–618, March 2011. 1

[41] Y. Sun, X. Chen, M. Rosato, and L. Yin. Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *IEEE SMCA*, 40(3):461–474, May 2010. 2

[42] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM TOG*, 36(4):95:1–95:13, July 2017. 10

[43] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. ACCV, pages 525–538, Berlin, Heidelberg. Springer-Verlag. 7

[44] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Niessner. Demo of face2face: Real-time face capture and reenactment of rgb videos. In *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH '16, pages 5:1–5:2, New York, NY, USA, 2016. ACM. 10

[45] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *FG*, pages 1–6, Sept 2008. 1, 2

[46] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *FG*, pages 211–216, Washington, DC, USA, 2006. IEEE Computer Society. 1

[47] S. Zafeiriou, M. Hansen, G. Atkinson, V. Argyriou, M. Petrou, M. Smith, and L. Smith. The photoface database. In *CVPR Workshops*, pages 132–139, June 2011. 2

[48] S. Zafeiriou and M. Pantic. Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter. In *CVPR Workshops*, pages 13–19, June 2011. 8, 9

[49] S. Zafeiriou, A. Tefas, and I. Pitas. The discriminant elastic graph matching algorithm applied to frontal face verification. *PR*, 40(10):2798 – 2810, 2007. 9

[50] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *IVC*, 32(10):692 – 706, 2014. 1, 2

[51] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, pages 3438–3446, June 2016. 1, 2

[52] C. Zhong, Z. Sun, and T. Tan. Robust 3d face recognition using learned visual codebook. In *CVPR*, pages 1–6, June 2007. 1